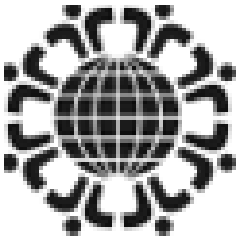
RESEARCH PLAN PROPOSAL

Design & Implementation of a Recommendation Model to Study Drug Target Interactions

For registration to the degree of
Doctor of Philosophy

IN THE FACULTY OF SCIENCE



IIS (DEEMED TO BE UNIVERSITY), JAIPUR

Submitted by

Harshita Bhargava
Enroll. No.: ICG/2018/26575

Under the Supervision of

Dr. Amita Sharma
Sr. Asst. Professor

Dr. Prashanth N Suravajhala
Senior Scientist, Systems Biology

Department of Computer Science & IT

August, 2019

1.Topic

Design & Implementation of a Recommendation Model to Study Drug Target Interactions

2. Introduction

The internet provides a variety of information in the form of research articles, news, audio and video content which can be searched using search engines or searching through the websites hosting them. The problem lies in the enormous content that may often confuse the user thereby resulting in irrelevant choices. Hence to address such problem, recommender systems (RS) were developed that automatically suggest the items of relevance according to the past preferences. Recommender systems (RS) are software tools that aim to provide items interesting to the users [Ricci *et al*. (2015)]. An item can be any object in the form of a good, a service or a process that an RS suggests to the user in an interaction through the internet. A user is generally an individual or a particular group of individuals who tend to use the recommender system for some specific purpose. The RS were defined as a means of assisting and imitating the process of using the suggestions of others to make fruitful choices as done in everyday life mostly in the form of verbal, textual messages or recommendation letters [Resnick *et al.*(1997)]. In the light of vast amount of choices available to the users, they tend to make wrong decisions. In order to relieve the users from the enormous amount of information while searching for a product or a service, RS act as an information filtering agent providing a limited but diversified and relevant subset of options. While RS have found applications in various domains ranging from e-commerce, e-learning, entertainment, travel and tourism, social networks, web search, gastronomy and health care [Lu *et al*. (2015)], in the recent past, a great attention has been attributed to the advent of Netflix challenge in the year 2006, won by the team Bellkor's Pragmatic Chaos [Koren *et al.*(2009)].

In the initial years the long tail phenomenon [Lee (2017)], especially in the case of e-commerce illustrated the need of implementing RS. In traditional commerce, the physical stores showcased only the popular items on an average from each category as opposed to the online stores, in e-commerce which weren't restricted to limited physical space. To cope up with these millions of product choices, online stores went for implementing RS which facilitated the task of filtering and showcasing not only popular but also unpopular or niche products from the long tail.

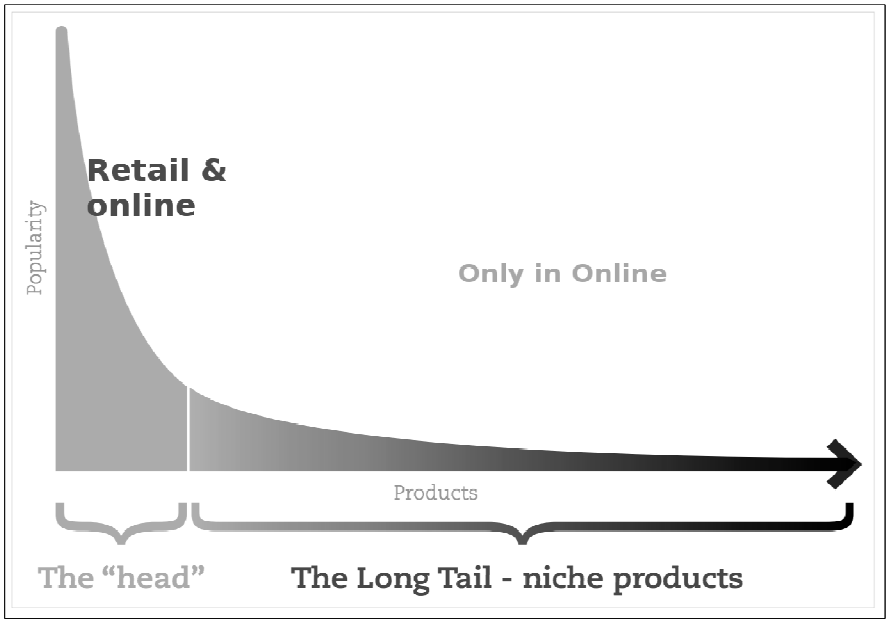


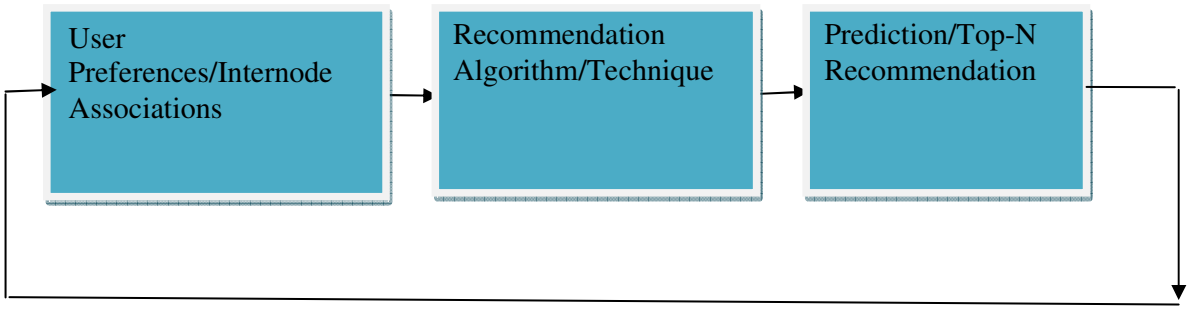
Figure (1):The Long Tail Phenomenon ('The Remarkable world of Recommender Systems – Towards Data Science', n.d.)

In a typical RS, the preferences of m users for n items are generally modelled as an m*n matrix ,also known as a preference, rating or utility matrix. The preference matrix is a sparse matrix with many blank entries and RS aims to predict these blank entries that model the preference of a user, using appropriate algorithms. The RS may also be modelled as a ranking problem as opposed to prediction depending on the domain in which it is applied. The user for which the items are recommended is referred as a target user. The users provide their feedback implicitly or explicitly. This implicit or explicit interaction from the past history of users plays a key role in recommending an item or a ranked list of items to the target user. In order to generate the relevant recommendations as an output, RS employ appropriate algorithms or techniques depending upon the objectives required to be achieved with respect to a particular domain. For example particularly in the e-commerce domain, although the primary objective is the annual profits earned, but then we are also concerned with the diversified, relevant and novel list of items that should be recommended to the users. This not only contributes towards higher profits but also increases the user trust and retention over years. RS differ in terms of application domain in which it is used and the objectives that need to be achieved but still a generic model describing the working of RS can be figured out.

2.1 Working Prototype of Recommendation System

The RS assist the users, decision makers or information seekers to select or choose from a subset of choices with the support of historical user preferences. The indicated preferences often take the form of explicit ratings or implicit ratings. Explicit ratings can be specified as binary ratings, continuous ratings, interval based ratings, ordinal ratings or unary ratings. In case of binary ratings one may specify only like or dislike as opposed to continuous or interval based ratings wherein a specific value (continuous value between a range or uniform interval based integer values) may be given to specify the level of preferences. In case of unary ratings the user can only specify the likings ie there is no way to specify the disliking. For example, the like button provided by Facebook is a mechanism to specify unary ratings. Implicit ratings are derived from the user actions such as user clicks, time spent on a web page while searching an item/information, items purchased etc and finally are transformed to unary ratings.

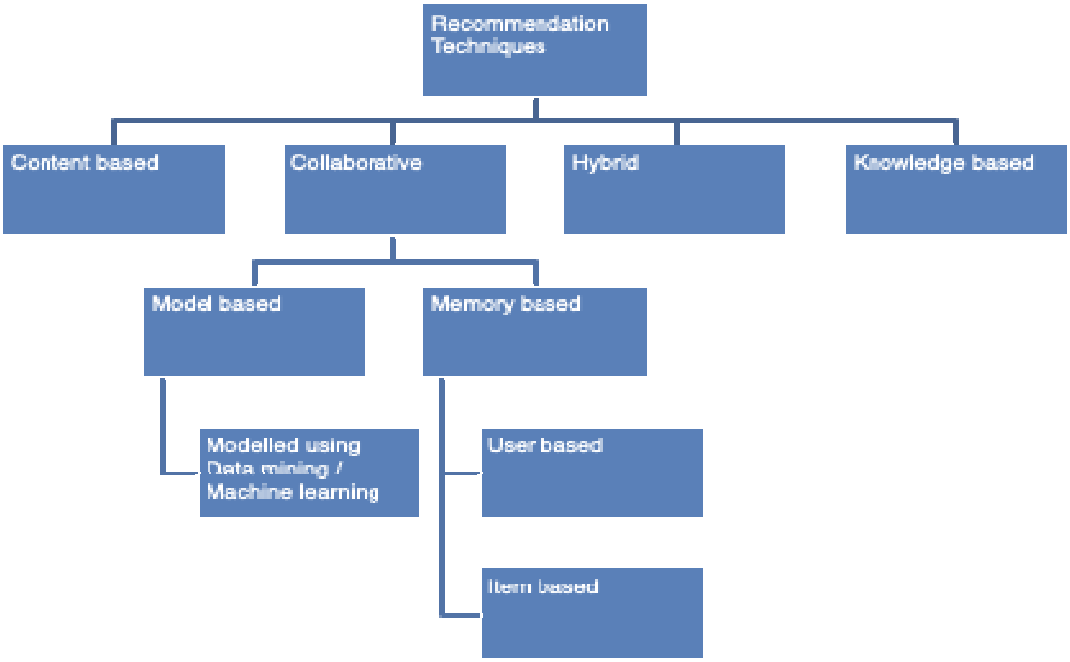
The ability of RS to predict the links in the large scale social networks has inspired many researchers to model their prediction problems as recommendation systems. They work by studying the existing underlying associations or interactions as links between entities that are represented using nodes in a network. The problem is reduced to predicting the future or tentative associations/interactions among these nodes. Hence the recommender techniques or algorithms can also be utilized in many problems in which prediction or ranked list of predictions is needed. For example [Natarajan *et al.* (2014)] used known gene disease associations to predict the unknown associations between disease and genes by utilizing matrix completion technique. A generic RS may contain three components: User preferences/Inter node associations, Recommendation algorithm/technique and prediction/Top-N Recommendation illustrated in Figure(2).



Figure(2):Recommendation components

The known user preferences or associations between nodes modelled as a network form the input for the recommendation algorithm. The recommendation algorithm/technique uses these further to predict the output in the form of predicted links in the network or ranked list of recommendations. The subsequent feedback after generating recommendations is again used to model the updated user preferences. The recommendation technique/algorithm is the central component of any RS which justify its applicability in a specific domain, are discussed in the next section.

2.2 Recommendation Techniques



Figure(3):Recommendation Techniques [Isinkaye *et al.* (2015)]

Recommender algorithms can be divided into two broad categories: personalized and non-personalized [Amin *et al.* (2019)], wherein personalized algorithms recommend different items depending on the target user while non-personalized algorithms suggest same set of items to each user ignoring the individual characteristics. Though the output of the recommendation process may differ depending upon the application domain where it has been implemented, the traditional recommendation techniques are collaborative, content based, hybrid and knowledge based filtering.

2.2.1 Collaborative filtering (CF) based techniques

Collaborative filtering (CF) techniques are based on human instinctive behaviour. The CF techniques assume that the users having similar tastes in the past are more likely to have the similar tastes in the future too. These techniques are based on the intuition of identifying the users similar to that of the target user X and using their ratings to predict the unknown ratings of the target user X. Conversely in order to recommend a target item P, the initial step is to determine a set of items I most similar to that of the target item P and then the ratings of items in set I specified by user X are used to predict the unknown ratings of user X. The methods employed in collaborative filtering are either memory based or model based methods [Ekstrand (2011)]. Memory based methods are also known as neighbourhood based collaborative filtering algorithms wherein the neighbourhood formation or the similarity calculation can be user based

or item based.In user based CF the similarity between target user and every other users is computed and the most similar users' ratings are used for generating recommendations.
In item based CF the similarity between the target item and every other item is computed and the most similar rated items by the user are used for generating recommendations. Item based CF as utilized by Amazon proves to be more scalable than user based CF when no of users increases since items tend to be smaller as compared to the users (Linden, Smith, & York, n.d.). So in case of item based CF similarity computation may be carried out offline to provide better scalability and comparable quality as with user based CF [Sarwar *et al.* (2001)]. In model based methods either data mining or machine learning model is used to learn the user preferences and generate predictions on the basis of the learned model. CF techniques suffer from the new user as well as new item problem which is also referred to as the cold start problem [Adomavicius *et al.* (2005)].

2.2.2 Content based (CB) techniques

In content based (CB) filtering method, the ratings and the buying history of the user is used to match the specific attributes of the items and the corresponding matches are the resultant recommended items. The recommendation depends upon the match between the rich user profile and the item specific attributes. For example if user X liked a science fiction movie Q then there is a high probability that he may like movies from the same genre in the future [Aggarwal (2016a)] This method suffers from the problem of overspecialization wherein the user may get the same kind of inherent and obvious recommendations lacking any surprisal element. This method is immune to cold start problem of new items as item attributes are utilised for computing the recommendations and are matched against the user profile. The limitation lies in handling cold start problem of new users wherein the user has no initial profile with preferences due to which the user may not receive appropriate recommendations as per his choice [Ricci *et al. (*2015)].The ability of these methods to explain the generated set of recommendations develops a sense of trust and transparency on the part of users [Aggarwal (2016a)].

2.2.3 Hybrid recommendation techniques

In order to overcome the demerits of the traditional recommendation techniques hybrid recommendation techniques have been proposed .Several recommendation techniques can be combined in order to overcome the challenges faced by each technique individually. According to [Burke (2007)] several hybridisation methods can be chosen such as:
Feature combination-uses a combination of features from different sources including ratings, content and context features for effective personalization.
Cascade-In this method the output of the first recommender technique is fine-tuned by the next technique to generate recommendations thereby reducing the recommendation list with precise and relevant items.
Feature augmentation-The output provided by one recommender system is used to create features to be used by the next. For example using a content based recommender we may impute the missing enteries by spurious ratings and then the entire rating matrix may be used by a collaborative recommender to predict the ratings. While computing similarity a proper weight assignment is required giving more weight to true ratings than the spurious ones.
Meta-level-The model build by the previous recommender is used by the latter. For example Fab, an RS build for an online news domain [Balabanović et al. (1997)] uses content based recommender to learn the user profile based on term vectors and then uses collaborative recommender to identify peer groups from the same user profiles though the final recommendation uses the initial rating matrix.
Switching-In this the algorithm switches between the recommenders and selects the one that produces the most accurate recommendation at that point of time ie may vary as per the specific situation.

Mixed-Here a mixed approach is used for recommending items having several subcomponents or in complex domains such as tourism wherein a single recommender may not perform well considering the different contexts. All recommendations from different techniques are provided to the user at the same time.
Weighted hybrid-In this the predicted ratings of several recommenders are combined to produce a single rating matrix using a weighted aggregation scheme.

2.2.4 Knowledge based (KB) techniques

Knowledge based techniques rely on the user's explicitly specified requirements in the form of user interaction with the system. These kind of recommendations are used when sufficient amount of ratings are not available or when the user is not specific about his requirements in the initial phase. They are useful with respect to items that aren't bought frequently such as expensive goods, financial policies, vehicles etc or when specific constraints drive the relevant set of choices for the intended users or group of users [Aggarwal (2016b)]. The most important advantage of using knowledge based recommendation is that it does not depend upon the user's ratings and hence the problem of cold start is eliminated. For example: In purchasing a house with specific no of rooms and specific plot area the knowledge based RS may impose the constraints on rooms and plot area attributes to generate recommendations. Knowledge based RS may involve case based or constraint based approaches [Aggarwal (2016b)]. In the case based approach the case is matched against the cases stored in the case base while in constraint based approach the user specified constraints are taken into consideration while generating the recommendations.

In particular all the above techniques have their pros and cons, so in order to select the appropriate technique we need to study the application domain and its characteristics in which the RS needs to be implemented. The collaborative filtering techniques, specifically memory based techniques suffer from the sparsity problem due to dependency on user ratings wherein many ratings are not specified in the explicit settings or are difficult to derive in the implicit settings. They are not scalable with respect to large datasets and also suffer from the new user and new item problem due to which similarity computation process does not include the users and items unless they have rated sufficient no of items or item has been rated by sufficient no of users respectively. The model based filtering needs to construct the model using machine learning or data mining prior to prediction but tend to address the problems of scalability and sparsity through clustering and dimensionality reduction techniques respectively [Su *et al.* (2009)]. The content based filtering technique suffer from the new user problem wherein the initial user profile has no items indicating his preferences to match against the item attributes and recommending the items/services. They also tend to overspecialize as per the learned user profile and are unable to generate much obvious recommendations. Knowledge based filtering techniques need to maintain a comprehensive knowledge base specifying rules to match against the explicitly specified user requirements which may be refined iteratively. The hybrid filtering techniques combine the merits of each of the techniques thereby providing accurate recommendations.

2.3 Applications of RS

The ability of RS to automatically infer the user preferences and suggest relevant information/products/services from a vast collection available online through an interactive interface have remarkably affected their development and implementation in several domains. The most prevalent application of RS is e-commerce wherein the RS tend to increase the cross-sell by suggesting additional related products with that in the cart and thereby increasing the sales and profits [Schafer *et al.*(1999)]. In the world of entertainment where the user has

several options including movies, music, streaming videos etc, RS act as sifters and provide relevant and top-n choices on TV, tablet or mobile devices. MovieLens and BookLens, developed as an expansion to GroupLens RS for recommending movies and books respectively were the earliest examples of recommender systems from the GroupLens research team ('GroupLens', n.d.) .Netflix [Gomez-Uribe *et al.* (2015)] ,initially developed as a DVD rental service of movies and television shows delivered through mail and later extended to streaming videos is a significant example of application of RS in the entertainment domain which uses an ensemble of RS algorithms for generating accurate and personalized recommendations. Social networking websites such as Facebook uses RS for predicting possible links in order to suggest friends in the social network. RS have considerable applications in gastronomy wherein they play a prominent role in suggesting food items in the form of recipes, meal plans or groceries [Trattner *et al.* (2017)]. RS applications also extend to research paper recommendations as in Mendeley, wherein research articles are recommended on a personalized basis [Vargas *et al.* (2016)]. The tourism websites such as TripAdvisor [Ricci (2014)] , also realized the benefits by implementing RS that suggest travel plans, hotels to individual users or groups having similar travel interests on a personalized basis and the users too share their travel feedback by sharing photos, reviews etc. The RS supporting e-tourism increase the level of trust among the users by providing better explanations. Since the medical field is also generating enormous patient data in the form patient electronic health/medical records (EHR/EMR) records, image data in the form of ultrasounds, x-rays, sensors etc typically referred as big data, medical RS tend to affect the lives of many people by exploring a set of people having similar conditions in the past as the current user's medical profile. This helps in exploring the similar cases and analysing their remedial steps taken up at at that particular point of time. Health RS have also gained popularity in the recent years that provide recommendations in the form of predicting risks for various diseases which tend to be comorbid, based on patient's health profile [Chawla *et al.*(2013)].

As evident from the stated applications, RS deal with enormous amount of data and needs to be scalable to handle the future needs with the same accuracy. In the age of Big Data where the data is not only characterised by its sheer volume and speed with which it is created but also by its nature which can be either structured, semi structured or unstructured. The ability to produce useful insights out of this big data still remains a challenge to the data scientists.One such source of big data results from the bioinformatics field in the form of DNA, RNA, protein sequences, microarray data, protein protein interaction data etc [Kashyap *et al.*(2016)].

2.4 Recommendation Systems in Bioinformatics

2.4.1 Introduction to Bioinformatics

Bioinformatics is an interdisciplinary field that develops and applies computational methods to analyse large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or discover new biology. The computational methods widely employ analytical methods, and mathematical modelling approaches besides machine learning and simulation (Nature, n.d.) .

Some of the most relevant terms used in bioinformatics are listed below:

- DNA stands for deoxyribonucleic acid, is a double stranded helical structured molecule containing biological instructions or genetic information that make each organism unique. The DNA is passed along with the genetic information from the parents to their offsprings during reproduction. A DNA molecule is made up of nucleotides which itself consists of three parts: a phosphate group ,sugar group and one of the four nitrogen bases that are A(adenine), C(Cytocine), T(Thymine) and G(Guanine).The order of these bases determine the instructions that they code for. DNA bases pair with each other where A always pairs with T and C always pairs with

G to form base pairs. In humans DNA resides in the nucleus of the cell called nuclear DNA while some DNA may form part of mitochondria called mitochondrial DNA. [National Human Genome Research Institute(NHGRI), 2015)]

- RNA stands for ribonucleic acid, is a single stranded molecule as opposed to DNA and consists of ribose sugar, phosphate group and the four nitrogen bases that are A(Adenine), C(Cytocine), G(Guanine) and U(Uracil). The base pairing is the same as in DNA except for A that pairs with U. RNA acts as an intermediary between DNA and proteins .The DNA is transcribed into RNA by a process called transcription and then RNA is translated into proteins by a process called translation. [Jones *et al*. (2004)]

- A gene is a basic unit of heredity made up of DNA .Every individual has two copies of each gene coming from each parent. They contain instructions for coding into proteins or other function so not all genes are involved in coding proteins. [Jones *et al*. (2004)]

- Proteins are large complex molecules that are required for structure, function and regulation of the body's tissues and organs. Each protein has unique a three dimensional structure and a specific function which is dependant on the sequence of amino acids. Proteins function in the form of antibodies, as structural components of body such as hair, muscle etc, as enzymes responsible for carrying out chemical reactions in cells, as messenger sending signals to other cells, organs and tissues. [Jones *et al*. (2004)]

- Drugs are the molecules used as medicines or as components in medicines to diagnose, cure, mitigate, treat, or prevent disease [Silverman *et al.* (2014)].

- Drug Targets-It is a molecule in the form of a protein in the body, associated with a particular disease process, acted upon by a drug to produce a desired therapeutic effect [Amaratunga *et al. (*2007)].

- Drug Repositioning- Drug repositioning is the identification of new uses of existing drugs to treat diseases other than the disease for which it was originally developed [Xue *et al. (*2018)].

2.4.2 Area of Computational Research

Big data is the inherent result of low cost computing technologies, but in order to harness value out of it organizations need to analyze it and draw inferences to support decision making. The heterogeneous nature of big data is attributed to a variety of sources including location data, image data, data generated through sensors, geospatial data, primary or secondary research. The availability of such heterogeneous data not only provides opportunities for data-driven decision making through data analytics but also poses serious challenges in terms of accuracy while maintaining the data privacy.

In the recent years, bioinformatics data has also grown at a very fast pace fuelled by the open source genetic databases and low cost and time saving next generation sequencing (NGS) or high throughput sequencing techniques [Pareek *et al.*(2011)]. The data sources maintained at EMBL's European Bioinformatics Institute only have 120 petabytes of data as of September 2017, which is expected to rise to exabytes scale by 2022 [Cook *et al. (*2018)] .The stored data in open source genetic databases can be used by the biologists, bioinformaticians, statisticians and computer programmers to develop hypotheses, draw conclusions and perform data assisted predictions.

The predictions can be in the form of drug target associations, side effects of drugs, drug to drug associations, protein to protein associations, gene disease associations etc. These predictions prove to be an important milestone in the process of either drug discovery or utilization of existing drugs for new diseases. The latter process of using existing pre-tested drugs for treating a disease other than that for which it was originally developed is better known as drug repositioning. The advantage of drug repositioning is in the reuse of existing drugs, eliminating the high cost of drug development and testing in the later phases. The problem of drug repositioning in bioinformatics research stems from the process of drug target interaction prediction as only after knowing new targets we may find new use of drugs.

Many computational approaches have been proposed to analyse this big data and produce inferences in a timely and cost effective manner as compared to the traditional wet lab experiments. Bioinformatics research has witnessed the use of RS algorithms in many applications ranging from drug repositioning, ranking genes for a specific disease, cancer drug response, personalised medicine, recommending drug combinations based on drug drug interactions, drug target interaction predictions to drug side effect predictions etc. In this research we consider the problem of drug target interaction prediction using known interactions or associations between drugs and targets with recommender algorithms handling the prediction task.

3. Review of Literature

The RS are software tools that act upon the known preferences/interactions/associations between entities thereby providing predictions or ranked list of predictions for the unknowns. In e-commerce applications, the entities are users and items/products, in gastronomy applications the entities are users and food items/groceries/menus/recipes so they basically act as an assistant for making reliable and accurate decisions from vast no of options available. The underlying recommender algorithm/technique plays a vital role in any of the application domain where RS has been implemented.

Amin *et al*. (2019) have discussed the CF techniques wherein the recommendations to a user rely on neighbourhood formed using similar users or items. The authors reviewed the techniques of collaborative filtering as: memory based and model based filtering along with the limitations of scalability, sparsity, and cold start issues. The scalability issue needs to be addressed with the growing no of users and items over time while sparsity refers to unavailability of preferences/ratings from most of the users. The cold start problem arises when either the user or item is new i.e. either the user hasn't rated any item or the item hasn't been rated by any user.
A similar contribution was made by Melville *et al*. (2017) who discussed the concept of content based, collaborative, as well as hybrid filtering techniques. The CB techniques are based on the similarity of the item attributes with that of the items in the user profile while hybrid techniques are based on the combination of the collaborative and content based techniques. They also elaborated the challenges of user cold start and serendipity with respect to content based techniques. Since the user is new so initially the user profile would be blank and would produce no recommendations to this user in a content based scenario referred as user cold start problem. As the user profile is matched against the item features or attributes so the content based RS tends to generate very obvious recommendations (non serendipitous) which isn't the case with collaborative techniques. Raghuwanshi *et al*. (2019) compared the nearest neighbour and latent factor model in CF based RS while outlining their goals including relevance, diversity, serendipity & novelty.Logesh *et al*. (2019) used the hybrid technique for supporting travel based application while Tarus *et al*. (2018) proposed a knowledge based filtering technique for e-learning using an ontology for representing knowledge.

The above discussed techniques have their pros and cons but in order to choose between them we need to study the application domain in which RS needs to be implemented with a set of desirable characteristics of accuracy, diversity, novelty, relevance etc. As far as big data is concerned medical data forms one of the most prominent source in the form of EHR/EMR data so to realize value out of this overwhelming amount of data significant work has also been done while implementing RS in the medical domain. Bao *et al.* (2016) proposed an intelligent recommender framework that recommends medicine/drug to an inexperienced doctor, medical student, intern or a patient based on the diagnosis data. A similar contribution was made by Chawla *et al.* (2013) by implementing a collaborative assessment and recommendation engine (CARE) for predicting disease risk based on similarity between the medical history of the patient and other patients. Hence RS attempt to address the big data generated in medical domain but the privacy concerns associated with EHR data and the risks involved limit their development as a common utility.

The field of bioinformatics characterized by similar characteristics of big data as in the medical domain has opened up new avenues for research facilitated by the availability and growth of open source biological databases. So in this sequence Hu *et al.*(2018) proposed a Top-N gene based collaborative filtering algorithm (GeneCF) that helps the doctors to find the list of genes for each patient that results in their effective diagnosis. The no of genes to be recommended is decided by an experimental design followed by a verification process from RNA expression of the diseased and healthy patients supported by cancer staining using Human protein atlas. Another research contribution by Zhang *et al.* (2015) was reported for predicting the unknown side effects of drugs from the known side effects using integrated neighbourhood recommender methods, Restricted Boltzmann machine machine and the ensemble of these methods. The problem of drug repositioning has also been addressed through the use of recommendation approaches as proposed by Ozsoy *et al.* (2018).Since the process of drug discovery, drug repositioning, side effect prediction typically relies on the drug target interaction prediction hence solving the underlying problem of DTI can prove beneficial in all these applications.

With regard to this Chen *et al.*(2018) classified drug target interaction prediction methods into ligand based, simulation docking based and chemogenomic based methods.

- Ligand based approach – predicts interaction for new ligand based on the known protein ligand but has a drawback that not all ligands are known for proteins.
- Simulation docking based approach – require 3 dimensional protein structures for simulation and have limited applicability because of the unavailability of 3d structures for many of the proteins.
- Chemogenomic based approach- Uses both the information related to drugs and targets at the same time. They are further classified as network based and machine learning based methods.

This paper majorly focused on the machine learning based methods and listed the available public databases containing only the interactions between the drugs and targets. The machine learning methods are further divided into supervised and semi supervised learning methods. Supervised learning methods are classified as

- Similarity based methods
- Feature based methods

The authors concluded that the DTI prediction using machine learning based methods can greatly accelerate the drug development process thereby limiting the scope for costly wet lab experiments. Though while considering DTI prediction task as a supervised learning problem, the lack of negative examples can produce inaccurate or biased results. So in order to handle these imbalanced datasets having greater no of unknown interactions than the known ones the authors proposed semi-supervised learning methods as a solution for DTI prediction. Ensemble ap-

proaches can be developed through multiple classifiers to improve the prediction results. As future work the authors suggested that the DTI prediction task needs to be modelled as a regression problem with datasets containing the continuous binding affinity scores as opposed to binary datasets.

A similar analysis reported by Nath *et al*. (2018) also focused on machine learning methods for DTI prediction giving a process flow for their application while listing the shortcomings. The process starts with the gathering of known drug target pairs and their respective selected features followed by training the machine learning algorithm. The process is then followed by k fold cross validation or LOOCV by splitting training and test set. As the last step the selection of the best prediction model is done. The authors concluded by listing the shortcomings of machine learning methods where most of the methods lack interpretability and produce biased results with the absence of experimentally verified non-interaction data for drug target pairs. With these indicated shortcomings, the problem of drug target interaction prediction from the known set of interactions can be formulated as a collaborative filtering problem from the recommender systems. Typically as in the case of social networks hitting a Like button is an indication of liking but there is no mechanism of showing the disliking. So not hitting the Like button may not necessarily mean lack of interest it may be attributed to lack of attention or time. The same may be compared with the drug target interaction data reported in databases in which the unknown drug target pairs can be attributed as either negative examples or unknowns may contain some positive examples as well. The solution to this could be adapted from the one class collaborative filtering algorithm defined for such kind of problems.

Wu *et al*. (2018) presented the network based methods for DTI prediction by modelling drugs and targets as nodes in a bipartite, tripartite or a heterogeneous network and the interactions as edges. Many of these methods were inspired from the recommender algorithms used in RS or the link prediction algorithms including NBI series methods, similarity inference methods, random walk based methods, local-community-paradigm methods and the simple path based methods. The advantage of these methods is that they do not need to consider the underlying 3D structures of targets and the non-interaction data of drugs and targets. Despite of these advantages these methods have evolved to predict DTI for known drugs as well as new drugs but fail to predict DTI for targets without known ligands. Also the network based methods neither consider the type of interaction nor the quantitative value of binding affinities. Though binding affinity was considered in EWNBI method but it resulted in poor performance as compared to the NBI method originally developed for RS.

Alaimo *et al*. (2016) discussed the use of recommendation technique in the DTI prediction and subsequently for drug repositioning. They modelled the drugs and targets as users and objects respectively as defined in a typical RS and the DTI network as a bipartite graph. They extended the basic NBI algorithm originally developed for RS by integrating the similarity information among targets and drugs respectively. The authors indicated that the RS algorithm has an advantage of predicting new drugs for a given target and new targets for a given drug at the same time. The limitation indicated exists with all the NBI based methods which require atleast one target associated with a drug and vice versa. The authors have proposed the use of DT-Hybrid in a tripartite graph for modelling the relation between drugs, targets and their associated diseases. An implementation of their algorithm in the form of a web application, DT-web is also discussed that performs drug target interaction and drug combination prediction.

The topological characteristics of the drug target network were explored by Lu *et al.* (2017) wherein the link prediction methods were used for DTI network modelled as a bipartite network with drugs and targets as nodes while the edges were represented with an association score from the MATADOR database. A comparative analysis between the proposed method and RBM from machine learning shows that approaches using a network topology tend to be accurate in the absence of prior information of drug and target related information. The disadvantage lies in the prediction of targets for new drugs or drugs for new targets.

Approaches used for Drug Target Interaction prediction

The overall problem of drug target interaction prediction is either modelled as a classification or regression learning problem. The experimentally verified associations are recorded in the biological databases which indicate the known interactions and unknown interactions using binary or continuous data in the form of an interaction matrix. The In silico approaches used to predict the drug target interaction are mainly divided into three categories namely ligand based molecular docking simulation based & chemogenomic approaches [Ezzat *et al.* 2018]. Among these categories the ligand based approaches use similarity of the query ligand with the known protein ligands for drug target prediction .These methods become inapplicable when only a small no of ligands are known for a target. Molecular docking based approaches involve the use of 3D structures of the targets for DTI prediction .The use of 3D structures of targets limits the use of such methods as for many of the targets such as membrane proteins, 3D structure is not known or is too complex to be derived. Chemogenomic approaches use both the information related to drugs and targets at the same time for inferring DTIs. Chemogenomic approaches can be further subdivided as network based and machine learning based methods [Chen *et al.* (2018)]. As per [Chen *et al*. (2018)] the machine learning based methods can be further classified as Bipartite local models, matrix factorisation models, deep learning models and semisupervised learning models.

4. Motivation

The objectives of an RS may differ with regard to design and evaluation of the corresponding recommender algorithm in a particular domain. However the RS algorithms have also been applied in bioinformatics for drug target interaction prediction and subsequently in drug repositioning [Alaimo *et al.*(2019)].The drug targets can be either DNA, RNA or proteins but proteins form the majority of known drug targets [Ezzat *et al.* (2018)]. The experimentally verified drug target interactions or association play a significant role in the workflow of drug discovery process as they are authentic, well tested and have lesser associated risks of drug development failures. At the same time, in silico or computational approaches for predicting drug target interactions, have gained popularity in the recent years due to publicly available data sources and the high cost involved both in terms of time, money and labour for conducting wet lab experiments. The use of computational methods helps in substantial reduction of the search space of candidate drug target pairs, for which in vivo or in vitro experiments may be carried out to assess their validity subsequently. The known interaction or association between the drugs and targets is recorded and stored in the biological databases or can also be extracted from medical texts using text mining techniques. In spite of storing the interactions among drugs and targets in the databases most of the interactions are unknown. As per the literature review these unknown interactions can either account for non-verified negative interactions or non-verified positive interactions. It is inherent from the literature review that either the method generates biased results due to lack of unreported negative or missing interactions or, are unable to predict drug targets for new drugs or new targets with no known interactions. So in order to solve the problem of drug target interaction prediction we may formulate the task as a recommender task from the domain of recommender systems wherein the users and items may be modelled by drugs & targets respectively.

The main objective of modelling drug target interactions (DTI) as a recommender task is the ability of recommender algorithms to deal with the sparse interaction data between drugs and targets. The idea would be to recommend a list of targets to the given drug. The design and corresponding evaluation of an appropriate recommendation model for accurate drug target prediction, even for existing as well as new drugs forms the basis for the research.

5. Objectives

1. To review existing prediction techniques for drug target interaction.

2. To study the application of recommendation techniques for drug target interaction based on the limitations identified in the existing techniques.

3. To identify the evaluation metrics for DTI recommendation.

4. To design or improve existing recommendation model(s) for predicting drug target interaction

5. To compare and evaluate the model with respect to other approaches.

6. To validate the model on a cohort of datasets.

6. Plan of work & Methodology
6.1 Work Plan

Phase 1 (6-8 months):

Literature survey of existing prediction techniques for drug target interaction.

Phase 2 (5-6 months):

Study the application of recommendation techniques for drug target interaction based on the limitations identified in the existing techniques.

Phase 3 (2-3 months):
Identify the evaluation metrics for DTI recommendation.

Phase 4 (6-8 months):

Design or improve existing recommendation model(s) for predicting drug target interaction.

Phase 5 (4-6 months):

Evaluation of the recommendation model with the set of identified metrics and a comparative analysis with other approaches.

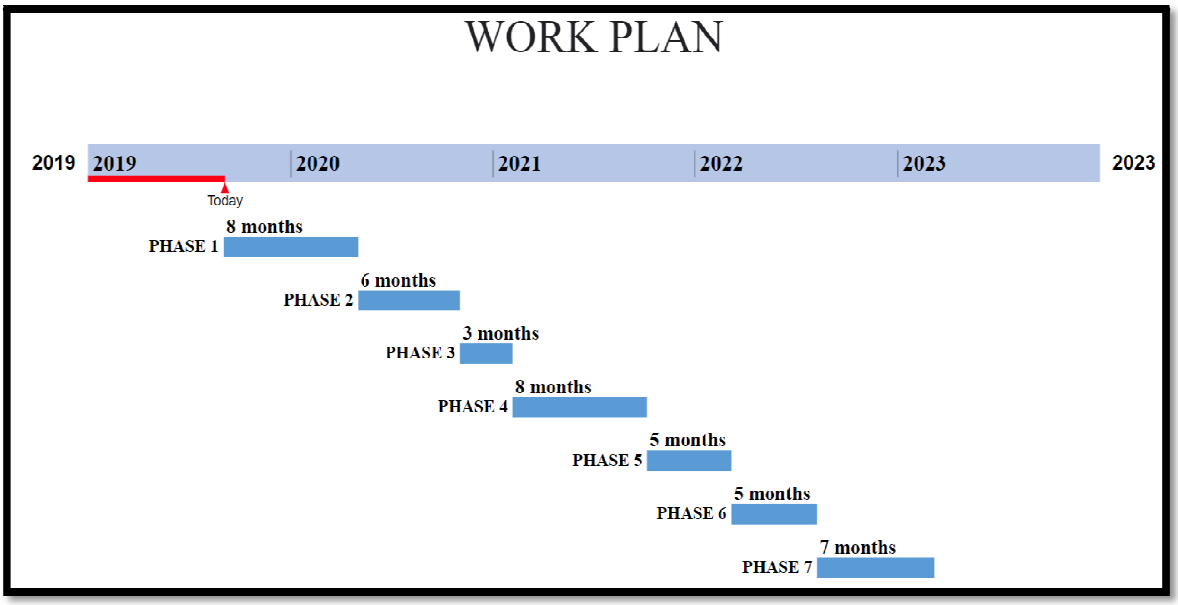
Phase 6(4-5 months):

Validate the model across a cohort of datasets.

Phase 7 (6-7 months):

Analyse the results and thesis writing.

6.2 Gantt Chart



6.3 Methodology

In order to attain the above mentioned objectives for DTI prediction using a recommendation model, the following methodology will be adopted:

- A study of existing prediction techniques used for DTI prediction will be done.
- The limitations of the existing techniques will be figured out.
- Evaluation metrics for DTI recommendation will be figured out.
- A new recommendation model or an improved model will be proposed for DTI prediction.
- Evaluation of the recommendation model will be done with appropriate set of metrics and will be compared with other related approaches.
- A validation of the recommendation model will be done across a cohort of datasets.
- Conclusions will be drawn on the basis of results obtained after evaluation through experiments.

7. Tools and Techniques

To carry out the proposed research we need tools for performing the various tasks.The following would be used for designing and evaluating the recommender model:
Python 3.7 (using any of the following packages)

- Surprise
- Crab
- Easyrec
- Lenskit

R 3.6

- DT-hybrid

Databases:

- DrugBank
- Binding MOAD
- MATADOR
- BRENDA
- KEGG
- SIDER
- ChEMBL

Dataset:

- Yamanishi *et al.* , 2008*(*http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/*)*

7. Facilities available at place of work

The research work will be undertaken in the Laboratories available in the university campus and the softwares and hardware provided will be used as a part of the research work.

8. References, Bibliography, Webliography

1. Adomavicius, G. and A. Tuzhilin (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.*IEEE TKDE*. **6**, 734-749.

2. Aggarwal, C. C. (2016a). Content-Based Recommender Systems, In Recommender Systems, pp. 139-166,Cham:Springer International Publishing.

3. Aggarwal, C. C. (2016b). Knowledge-Based Recommender Systems, In *Recommender Systems,* pp.167-197.,Cham: Springer International Publishing.

4. Alaimo S., Giugno R., Pulvirenti A. (2016) Recommendation Techniques for Drug–Target Interaction Prediction and Drug Repositioning. In *Data Mining Techniques for the Life Sciences (*Edited by O.Carugo, F. Eisenhaber), pp. 441-462, vol 1415. Humana Press, New York, NY.

5. Alaimo, S., and A. Pulvirenti (2019). Network-Based Drug Repositioning: Approaches, Resources, and Research Directions. In *Computational Methods for Drug Repurposing* (Edited by Q. Vanhaelen),**1903,** pp. 97–113, Humana Press, New York, NY.

6. Amaratunga, D., H. Göhlmann, and P. J. Peeters (2007). Microarrays. *Comprehensive Medicinal Chemistry II*, **3**,87–106.

7. Amin S.A., J. Philips, N. Tabrizi (2019) .Current Trends in Collaborative Filtering Recommendation Systems. In *Services – SERVICES 2019*.(Edited by Y. Xia, LJ. Zhang), Lecture Notes in Computer Science, **11517**,pp. 46-60 ,Springer, Cham.

8. Balabanović, M., and Y. Shoham (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, ***40***(3), 66–72.

9. Bao, Y., and X. Jiang (2016). An intelligent medicine recommender system framework. Proceedings of the 2016 IEEE 11th Conference on Industrial Electronics and Applications, ICIEA 2016, 1383–1388.

10. Burke, R. (2007). Hybrid web recommender systems. In *The adaptive web*,pp. 377-408, Springer, Berlin, Heidelberg.

11. Chawla, N. V. and D. A. Davis(2013). Bringing big data to personalized healthcare: a patient-centered framework. *Journal of General Internal Medicine*, ***28****(3)*, 660-665.

12. Chen, R.,X. Li, S. Jin, J. Lin and J. Liu(2018). Machine Learning for Drug-Target Interaction Prediction. *Molecules*, ***23***(9), 2208.

13. Cook, C. E., M. T. Bergman, G. Cochrane, R. Apweiler and E. Birney(2018). The European Bioinformatics Institute in 2017: data coordination and integration. *Nucleic Acids Research*, ***46***(D1), D21–D29.

14. Ekstrand, M. D. (2011). Collaborative Filtering Recommender Systems. *Foundations and Trends® in Human–Computer Interaction*, ***4***(2), 81–173.

15. Ezzat, A., M. Wu, X. Li, L. and C.K. Kwoh (2018). Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. Brief Bioinformatics, **8**,1-21.

16. Gomez-Uribe, A. Carlos and N. Hunt (2015). The Netflix Recommender System. *ACM Trans. Manage. Inf. Syst.* ***6***(4), 1–19.

17. GroupLens. (n.d.). Retrieved 30 June 2019, from https://grouplens.org/

18. Hu, J., S. Sharma, Z. Gao and V. Chang (2018). Gene-based Collaborative Filtering using recommender system. *Computers and Electrical Engineering*, ***65***, 332–341.

19. Isinkaye, F. O., Y. O.Folajimi and B. A. Ojokoh(2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, ***16***(3), 261–273.

20. Jones, N. C., P. A. Pevzner and P. Pevzner (2004). *An introduction to bioinformatics algorithms*. MIT press.

21. Kashyap, H., H. A.Ahmed, N.Hoque, S.Roy and D. K. Bhattacharyya (2016). Big data analytics in bioinformatics: architectures, techniques, tools and issues. *Network Modeling Analysis in Health Informatics and Bioinformatics*, ***5***(1).

22. Koren, Y., R.Bell and C. Volinsky (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, ***42***(8), 30–37.

23. Lee, J. (2017). Recommendation systems. In *Big Data and Computational Intelligence in Networking,* pp. 227–264, CRC Press.

24. Linden, G., Smith, B., & York, J. (n.d.). *Amazon.com Recommendations*. IEEE Computer Society. Retrieved from http://computer.org/internet/

25. Lu, J., D. Wu, M. Mao, W. Wang and G. Zhang (2015). Recommender system application developments: A survey. *Decision Support Systems*, ***74***, 12–32.

26. Lu, Y., Y. Guo and A. Korhonen (2017). Link prediction in drug-target interactions network using similarity indices. *BMC Bioinformatics*, ***18***(1),39.

27. Melville, P. and V. Sindhwani(2017). Recommender Systems. In *Encyclopedia of Machine Learning and Data Mining ,* pp. 1056–1066 ,Boston, MA: Springer US.

28. Nath A., P. Kumari and R. Chaube (2018).Prediction of Human Drug Targets and Their Interactions Using Machine Learning Methods: Current and Future Perspectives. In *Computational Drug Discovery and Design*(Edited by M. Gore, U. Jagtap),pp. 21-30, Humana Press, New York, NY.

29. Natarajan, N. and I. S. Dhillon (2014). Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*, ***30***(12), 60–68.

30. National Human Genome Research Institute. (2015). Deoxyribonucleic Acid (DNA) Fact Sheet. Retrieved 4 July 2019, from https://www.genome.gov/25520880/deoxyribonucleic-acid-dna-fact-sheet/

31. Nature. (n.d.). Computational biology and bioinformatics - Latest research and news | Nature. Retrieved 2 July 2019, from https://www.nature.com/subjects/computational-biology-and-bioinformatics

32. Ozsoy,M.G., Özyer, F.Polat and R.Alhajj (2018). Realizing drug repositioning by adapting a recommendation system to handle the process.

33. Pareek, C. S., R. Smoczynski and A. Tretyn (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, ***52(***4), 413–435.

34. Raghuwanshi S.K., R.K. Pateriya (2019) Collaborative Filtering Techniques in Recommendation Systems. In *Data, Engineering and Applications* (Edited by Shukla R., J. Agrawal, S. Sharma, G. Singh Tomer, pp. 11-21, Springer, Singapore.

35. Resnick, P. and H. R.Varian(1997). Recommender systems. *Communications of the ACM*, ***40***(3), 56–58.

36. Ricci, F. (2014). Recommender Systems: Do They Have a Viable Business Model in e-Tourism?.

37. Ricci, F., B. Shapira and L.Rokach (2015). Recommender systems: Introduction and challenges. In *Recommender Systems Handbook, Second Edition,* pp.1–34, Springer US, Boston.

38. Sarwar, B., G. Karypis, J. Konstan and J. Reidl (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the tenth international conference on World Wide Web - WWW '01*,pp. 285–295, ACM Press, New York ,USA.

39. Schafer, J. Ben, J. Konstan and J. Riedi(1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce - EC '99*,pp. 158–166, ACM Press, New York, USA.

40. Silverman, R. B., andM. W.Holladay (2014). Introduction. In *The Organic Chemistry of Drug Design and Drug Action,* pp. 1–17,Academic Press.

41. Su, X. and T. M. Khoshgoftaar (2009). A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*,**2009**, 1–19.

42. The Remarkable world of Recommender Systems – Towards Data Science. (n.d.). Retrieved 28 June 2019, from https://towardsdatascience.com/the-remarkable-world-of-recommender-systems-bff4b9cbe6a7

43. Trattner, C. and D. Elsweiler (2017). Food Recommender Systems: Important Contributions, Challenges and Future Research Directions. Retrieved from http://arxiv.org/abs/1711.02760

44. Vargas, S., M. Hristakeva and K. Jack (2016). Mendeley. In *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16* pp. 365–365, ACM Press, New York, USA.

45. Wu, Z.,W.Li, G.Liu, and Y. Tang (2018). Network-based methods for prediction of drug-target interactions. Front Pharmacol ,**9,**1134**.**

46. Xue, H., J. Li, H. Xie and Y. Wang(2018). Review of Drug Repositioning Approaches and Resources. *International Journal of Biological Sciences*, ***14***(10), 1232–1244.

47. Zhang, W., H. Zou., L. Luo, Q. Liu, W. Wu and W. Xiao (2015). Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing, **173**, 979–987.